Xử Lý Ngôn Ngữ Tự Nhiên

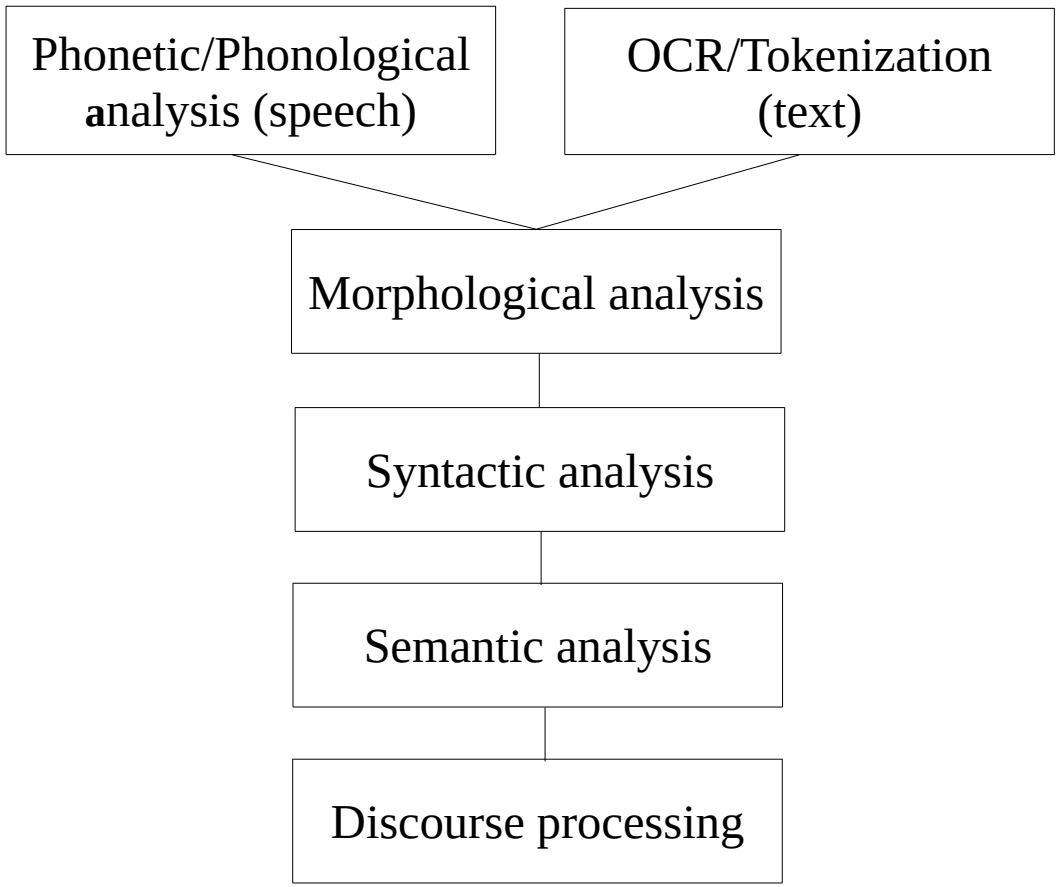
Giới Thiệu

Quy Nguyen

Nội dung

Giới thiệu xử lý ngôn ngữ tự nhiên

- Mục tiêu của NLP: Thiết kế các thuật toán để máy tính có thể hiểu ngôn ngữ tự nhiên
- Các mức độ của bài toán NLP



Knowledge of language

A complex language behavior requires various kinds of knowledge of language:

- Phonetics and Phonology — knowledge about linguistic sounds
- Morphology — knowledge of the meaningful components of words
- Syntax — knowledge of the structural relationships between words
- Semantics — knowledge of meaning
- Pragmatics — knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse — knowledge about linguistic units larger than a single utterance

Morphological analysis

- Text Normalization, such as stemming and lemmatization: The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
  - **Stemming** algorithms usually refer to a *heuristic process*, work by cutting off the end or the beginning of the word.
    studies -> studi
    studying -> study
  - **Lemmatization**, on the other hand, takes into consideration the morphological analysis of the words. *studies -> study*
    studying -> study
  - Text normalization is frequently used when converting text to speech, used for searching, etc.
- Tokenization, such as word segmentation in Vietnamese

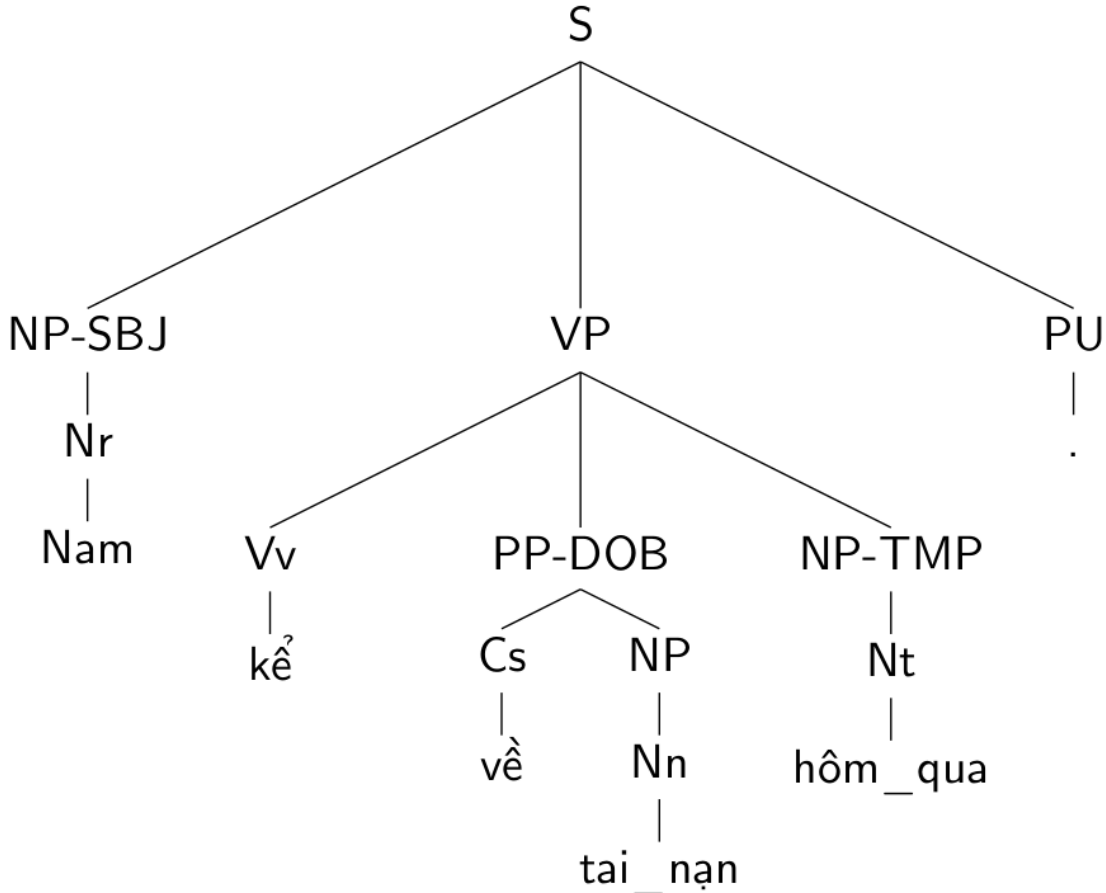
Morphological analysis

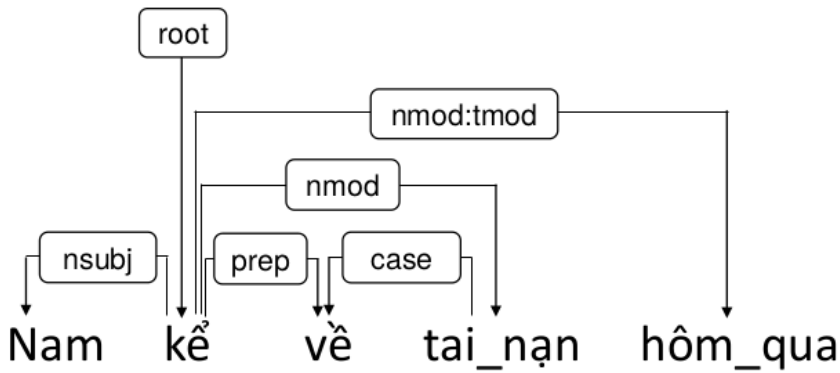
- Ứng dụng trong các hệ thống hỏi đáp: HAL is capable of producing contractions like *I'm* and *can't*. Producing and recognizing these and other variations of individual words (e.g., recognizing that doors is plural) requires knowledge about morphology, the way words break down in to component parts that carry meanings like singular versus plural

Morphological analysis

- POS tagging:
  VNese sentence: *Nam kể về tai nạn hôm qua.*
  WS: *Nam kể về tai_nạn hôm_qua .*
  POS tagging: *Nam/Nr kể/Vv về/Cs tai_nạn/Nn hôm_qua/Nt ./PU*

Syntactic analysis





Syntactic analysis

- by-phrases:
  - *How much Chinese silk was exported to Western Europe* **by the end of the 18th century***?*
  - *How much Chinese silk was exported to Western Europe* **by southern merchants***?*
- The sequence of words do not make sense:
  - *I'm I do, sorry that afraid Dave I'm can't.*

Semantic analysis

- The work of semantic analyzer is to check the text for meaningfulness.
- Semantics refers to the meaning of words in a language and the meaning within the sentence.
- Semantics considers the **meaning of the sentence without the context**.
- Semantics is just the meaning that the grammar and vocabulary impart, it does not account for any implied meaning.

Ex: *How much Chinese* **silk** *was* **exported** *to Western Europe by the* **end** *of the 18th century?*

Pragmatics vs discourse

- Pragmatics studies the ways **in which context contributes to meaning**
- Discourse analysis looks how the sentences are glued together

Ex:

- Speaker 1: *would you like to go for a drink?*
- Speaker 2: *great! what time?*

The discourse analysis would be looking at how this communicative event works mechanically whereas the pragmatist would be looking at underlying (implicit) meanings (in this case 'drink' means trip to the pub, for example).

Coreference resolution

Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.
Ex:

- *How many states were in the United States that year?*

Một vài ứng dụng NLP

- Kiểm tra chính tả, tìm từ khóa, tìm từ đồng nghĩa
- Rút trích thông tin từ website, chẳng hạn như giá sản phẩm, thời gian, địa điểm, tên người, tên tổ chức
- Phân loại văn bản
- Dịch máy
- Hỏi - đáp

Ứng dụng NLP trong công nghiệp

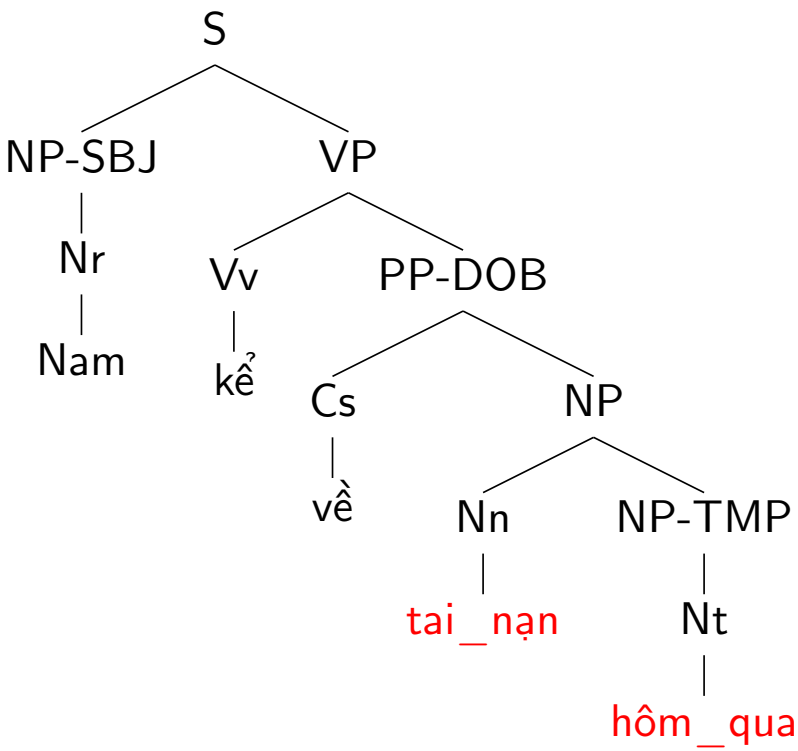
- Tìm kiếm
- Quảng cáo online
- Dịch tự động
- Phân tích cảm xúc khách hàng
- Nhận dạng tiếng nói
- Hỗ trợ khách hàng tự động

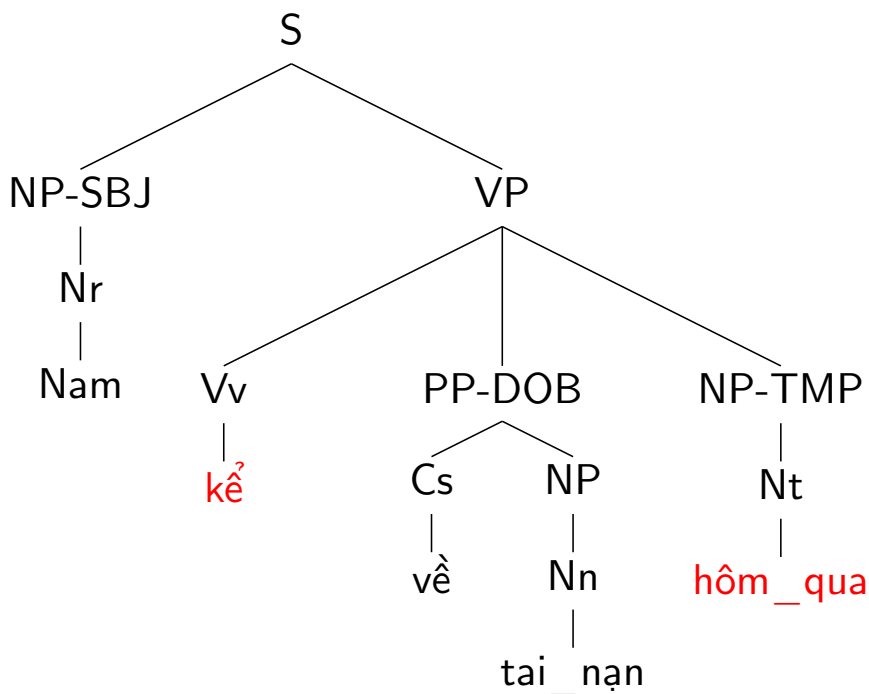
Tại sao NLP khó?

- Sự phức tạp/nhập nhằng trong việc sử dụng ngôn ngữ
- Ví dụ:
  - Đổi trật tự từ có thể tạo ra câu có nghĩa khác nhau:
    - *Con mèo này trắng*
    - *Con mèo trắng này*
  - Đảo trật tự từ không làm thay đổi nghĩa của câu:
    - *Khuyết điểm này tôi đang cố khắc phục.*
    - *Tôi đang cố khắc phục khuyết điểm này.*
  - Một câu có thể hiểu theo nhiều nghĩa
    - *Nam kể về tai nạn hôm qua.*
    - *Nam kể về tai nạn hôm qua.*
    - *Cô ấy đang bơi. {to swim}*
    - *Cô ấy đang bơi. {to struggle with one's job}*

Disambiguate

- Phân tích cú pháp





Disambiguate

- Part-of-speech tagging:
  - *Cổ/Pp đẹp quá.*
    $\rightarrow$ cổ = she
  - *Cổ/Nn dài quá.*
    $\rightarrow$ cổ = neck
- Word sense disambiguation:
  - Tin học: *memory* = *bộ nhớ*
  - Y khoa: *memory* = *trí nhớ*

Yêu cầu

Có kiến thức về:

- Python
- Giải tích, đại số tuyến tính, xác suất thống kê
- Máy học

Xử lý ngôn ngữ tự nhiên và các vấn đề liên quan

- Xử lý ngôn ngữ tự nhiên là sự giao thoa giữa:
  - Ngôn ngữ học (linguistics)
  - Máy học (machine learning)
  - Trí tuệ nhân tạo (AI)
  - Khoa học máy tính (computer science)

Ngôn ngữ học

- Mục tiêu của ngôn ngữ học: Tìm hiểu cách thức ngôn ngữ được vận hành, ví dụ như:
  - Ngôn ngữ thuộc họ nào, loại hình nào, các họ ngôn ngữ có liên quan với nhau như thế nào?
  - Tiêu chí để xác định một câu đúng ngữ pháp là gì? Chúng ta có thể áp dụng các tiêu chí này trên nhiều ngôn ngữ khác nhau hay không?
  - Ngôn ngữ thay đổi như thế nào và tại sao phải thay đổi?
  - Con người học ngôn ngữ đầu tiên của họ như thế nào và có gì khác khi họ học ngôn ngữ thứ 2?
- NLP tận dụng sự hiểu biết từ ngôn ngữ học để xây dựng các kỹ thuật ngôn ngữ.

Ảnh hưởng của ngôn ngữ học đối với NLP

- Tiếng Việt thuộc ngữ hệ Đông Nam Á (Austroasiatic languages), nhánh Mon-Khmer, loại hình đơn lập
- Tiếng Anh thuộc ngữ hệ Ấn-Âu (Indo-European languages), loại hình hòa kết
- Đặc điểm ngôn ngữ ảnh hưởng việc xử lý ngôn ngữ, ví dụ
  - Tiếng Anh có biến cách, từ được phân biệt bằng khoảng trắng
    - Stemming, lemmatization
    - POS tags được thiết kế dựa trên word forms, ví dụ:
      PRP: Personal pronoun (I, you, ...)
      PRP$: Possessive pronoun (mine, yours, ...)
  - Tiếng Việt không có biến cách, từ không được phân biệt bởi khoảng trắng
    - Word segmentation
    - POS tags được thiết kế dựa trên khả năng nết hợp và chức năng cú pháp của từ, ví dụ:

Máy học

- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1]
  - Hầu hết các ứng dụng NLP đạt kết quả tốt hiện nay đều sử dụng máy học bởi vì ngôn ngữ rất phức tạp
  - Ví dụ: Hệ thống dịch máy được xây dựng dựa trên kho ngữ liệu song ngữ, không phải dựa trên luật hay từ điển.
- Vì vậy, có nền tảng về máy học là yêu câu thiết yếu của khóa học NLP

[1] https://en.wikipedia.org/wiki/Machine_learning

Trí tuệ nhân tạo (AI)

- In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans.[2]
- Mục tiêu của AI là tự động hóa tiềm năng trí tuệ của con người
  - Ngôn ngữ là một khía cạnh nền tảng của trí thông minh con người
  - Ngôn ngữ có ích trong việc giải quyết các nút thắt về tri thức, cung cấp kiến thức cho hệ thống AI để nó có thể đưa ra những kết luận hữu ích
  - Suy luận cũng cần thiết cho việc hiểu ngôn ngữ
    a. The trophy doesn't fit in the suitcase because it is too big.
    b. The trophy doesn't fit in the suitcase because it is too small.

[2] https://en.wikipedia.org/wiki/Artificial_intelligence

Khoa học máy tính

- "A computer scientist studies the theory of computation and the practice of designing software systems"[3]
  - "The theory of computation is the branch that deals with how efficiently problems can be solved on a model of computation, using an algorithm."

[3] https://en.wikipedia.org/wiki/Computer_science

Khoa học máy tính

- NLP thể hiện một vài khía cạnh cốt lõi của khoa học máy tính
  - Ngôn ngữ tự nhiên có thể được mô hình hóa sử dụng lý thuyết ngôn ngữ hình thức (formal language theory), tương tự như lý thuyết được sử dụng để phân tích ngôn ngữ lập trình
    Ex: *Let* $\sum = \{a, b, c, \ldots, y, z\}$. *Then* $\sum^*$ *is the set of all strings over the Latin alphabet.*
  - Dữ liệu ngôn ngữ tự nhiên cũng yêu cầu các thuật toán hiệu quả, có thể phân thích độ phức tạp về không gian và thời gian thực thi.
  - Các thuật toán này phải được thực thi trên các kiến trúc đa dạng, chẳng hạn như các hệ thống phân tán, GPU, hoặc các thiết bị mobile.